Survey of Gmail Spam Detection and Classification Existing Approaches

Javed Akhtar Khan
Dept of Computer Science & Engineering
Indur Institute of Technology

Abstract- *Spam detection and its filtering techniques survey is aim of this paper. Spam and its detection is not simple task for achieving this task many researcher are introduce so many techniques ,algorithm , mining approach , frame work ,mining concept , automatic detection of spam , various parameter etc . This paper is include the analysis table based on the literature study for Spam Detection and its classification.*

Keywords- Gmail ,Inbox , Mail Cluster , Spam, Spam Classification

I. INTRODUCTION

Literature survey

Ref[1]- In this paper author are describe how a large webmail service uses reputation to classify authenticated sending domains as either spammy or not spammy. For this work Both SPF and Domain Key authentication are used to identify who the sender of the mail . in this paper author describe a simple formula for how we calculate the reputation and how it is used to classify incoming mail. In this paper show in general how domains, both good and bad, get distributed among the various reputation values, and also show the reputation values for a few specific domains. In this paper researcher also describe some of the problems associated with this reputation system, and propose some recommendations for senders to avoid those problems.

Ref[2]- This is a survey paper based on some popular classification technique to identify whether an email is spam and non-spam. For representing spam mails , in this paper author are use vector space model(VSM).In this paper author are used the Naïve Bayesian Classification , Decision Trees and K-Nearest Neighbor Algorithm used .

Ref[3]-In this paper researcher are proposed the technique for identification of malicious Emails form the inbox .In this research paper author are work with the mass created emails , and work for the ,Unauthorized email systems are not registered in server. The un-authenticated email systems will forward the unwanted mails with a waste content and come into inbox and stored. These emails are dangerous. These emails called as a spam or phishing emails. In this paper author are prospering and filter the spam or phishing emails with different kinds of techniques. For this work researcher are make a new method called Procedure of Tailored , this method is based on the detection of characteristics pattern matching. In this paper author also work for detection of persistent and recipient based attacks. Repeated intrusion attack attempts are the type of under persistent attacks. Identifies the list of bad senders and display the recipient oriented emails. The purpose of this paper is to identify the Detecting targeted malicious email in the mass generated emails. Here filter the spam or phishing emails with different kinds of techniques. In this paper author are proposed the decision-tree classification algorithms split each node using the best split from all available features. The best split is that which provides the most separation in the data. In this paper the Proposed System are as Using new detection and filtering techniques starts the detection of phishing and spam emails. Every mail verifies with probability distribution characteristics Once all the characteristics are satisfied in Meta data structure then to allow in inbox. In inbox all the mails are placed here only trust worthy mails of content. Meta data structure follows some data mining techniques. In total number of mails after performing the preprocessing operation then to apply here classification techniques. Classification gives the results like recipient and persistent mails. The main advantages of this work is : It can display bad list attacker's information. It is reduced the energy levels and cost. In this research work author are describe the following module these are Targeted malicious attacks structure, Dataset construction, , Targeted Malicious email, Non targeted malicious email Persistent, Recipient oriented attacks detection.

Ref[4]-This is survey paper this paper consists of comprehensive study of spam detection algorithms under the category of content based filtering and rule based filtering. The implemented results have been benchmarked to analyze how accurately they have been classified into their original categories of spam and ham. Further, a new filter has been suggested in the proposed work by the interfacing of rule based filtering followed by content based filtering for more efficient results.

Ref[5]-In this work, we consider a system that automatically suggests relevant view filters to the user for the currently viewed messages. In this paper author are propose several ranking algorithms for suggesting useful filters. In this paper researcher are work for suggesting that such systems quickly filter groups of inbox messages and find messages more easily during search. The main objective of this paper is to generate the automatically generating a list of view filters relevant to the displayed messages. Researcher are make a filters are implemented as searches, such as a search for all messages in the inbox from a discussion list. After that author are call task Search Operator Suggestion, where search operators are special terms that retrieve emails based on message metadata, author are build a mail filter system for Gmail (Google Mail) using search operators and develop several search operator rankers using features of the user, mailbox and machine learning. When the present the user interface and system for collecting usage data. Several ranking systems and features are evaluated on collected data; our rankers yield high quality suggestions.

Ref[6]-This paper is also work for the spam detecting process and spam filtering technique .The novel approach is in this paper is its work offline data collection based on the some spam filtering algorithm author make new email system by combining the concepts of Bayesian Spam Filtering Algorithm, Iterative Dichotmiser 3(ID3) Algorithm and Bloom Filter. A weighted email attribute based classification is proposed to mainly focus to encounter the issues in normal email system. This proposal is implemented using .net and sample user-Id for knowledge base.A weighted email attribute based classification is proposed to mainly focus to encounter the issues in normal email system. It makes the user to feel more securable by means of detecting and classifying such malicious mails when the user checks the inbox by notifying with different colors for spam (red), suspected (blue) and ham (green) mails. These type of classification helps to formulate an effective utilization of our email system.

Ref[7]- Now Again this paper is also include the Spam Classification analysis using the data mining Tool , in this paper author using the Supervised Learning Approach. This paper explores and identifies the use of different learning algorithms for classifying spam messages from e-mail. A comparative analysis among the algorithms has also been presented.

Ref[8]- This paper is based on the Spam Mail Detection using the data mining and show the comparative analysis In this study, author analyze various data mining approach to spam dataset in order to find out the best classifier for email classification. In this paper researcher analyze the performance of various classifiers with feature selection algorithm and without feature selection algorithm. Initially researcher are do experiment with the entire dataset without selecting the features and apply classifiers one by one and check the results. After that apply Best-First feature selection algorithm in order to select the desired features and then apply various classifiers for classification. In this study it has been found that results are improved in terms of accuracy when we embed feature selection process in the experiment. Finally author are found Random Tree as best classifier for spam mail classification with accuracy = 99.72%. Still none of the algorithm achieves 100% accuracy in classifying spam emails but Random Tree is very nearby to that.

Ref[9]- This paper is work for the spam filtering in this Paper describe solutions and analyze the bandwidth of the network by removing the unknown user from the email spammers lists. In this paper author are make a new approach that is called the "User unknown" within the SMTP-dialogue. This approach is only useful if the error message is generated in the SMTP dialogue, i.e.as soon as the sending party sent the "RCPT TO:"-line, the server should respond with the SMTP status code 550 "User unknown".

Analysis Table -

Ref No	Objective of Paper	Applied Approach	Algorithm used	Model
Ref[1]	This show the authentication approach for the large web mail service main objective of this paper is reputation of mail and identified the who sender .	SPF and Domain Key authentication are used to identify	NO this work is based on the IP address and Mail Address reputation .	NO
Ref[2]	This is survey paper based on the classification of spam mail and non spam mail in the inbox.	---	Native Bayesian Algorithm , Decision Trees algorithm and K-Nearest Neighbor Algorithm	Vector Space Model VSM used
Ref[3]	This paper work for the detection of persistent and recipient based attack , and filter the un-authenticated email objective of this paper is detecting the targeted malicious email .	A Meta data structure , and also display the bad list attacker; a new technique are proposed by the author for detecting the phishing and spam email , method called Procedure of Tailored	Decision Tree Classification Algorithm are used	No
Ref[4]	This paper is only include the study of various Spam detection Algorithm	This is survey paper and conclude the Rule based filtering approach is more efficient over the content based approach	NO only compare the existing spam detection Algorithm	NO
Ref[5]	This work is based on the filtering the viewed massages . In this paper researcher are work for suggesting that such systems quickly filter groups of inbox messages and find messages more easily during search.	Automatically Generating a list of view filters relevant to the displayed messages . Search Operator Suggestion, where search operators are special terms that retrieve emails based on message metadata,	Ranking Algorithm are used and make a mail filter system	

Ref[6]	Spam filtering problem , weighted email attribute based classification is proposed	A new Email system are introduce by the author that is combination of Bayesian Spam Filtering Algorithm, Iterative dichotomies 3(ID3) Algorithm and Bloom Filter. A color combination are make easy to filter the Email from the Spam ,Suspected	Combination of Bayesian Spam Filtering Algorithm, Iterative Dichotomies 3(ID3) Algorithm and Bloom Filter.	NO
Ref[7]	This paper is based on the Spam classification and proposed the comparative analysis among the some classification algorithms	This paper include the Concept of Supervised Learning Approach.	Comparative Analysis of Algorithm	No
Ref[8]	In this study, author analyze various data mining approach to spam dataset in order to find out the best classifier for email classification. In this paper researcher analyze the performance of various classifiers with feature selection algorithm and without feature selection algorithm	The entire dataset without selecting the features and apply classifiers this is the in concept of this paper , After that researcher are apply the Selection Algorithm	Best-First feature selection algorithm	
Ref[9]	Spam filtering using the bandwidth of network by removing the unknown user from the email spammers list .	With the help of network bandwidth researcher are filter and Spam form the inbox ,in this paper author are introduce the user unknown approach	NO	NO

REFERENCE

[1] Ref[1]- Taylor et al “Sender Reputation in a Large Webmail Service “Bradley Taylor Google Inc. 1600 Amphitheatre Parkway Mountain View, CA 94043

[2] Ref[2]- Ankita et al “Survey Paper on Effective Email Classification into Spam and Non-Spam Mails “International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 4, April 2015 ISSN(Online): 2320-9801 ISSN (Print): 2320-9798.

[3] Ref[3]- M. Eswra et al “Identifying Malicious Email” IJSRE ISSN - 2250-1991 Volume : 3 | Issue : 7 | July 2014

[4] Ref[4]-Sahil et al “COMPARISON AND ANALYSIS OF SPAM DETECTION ALGORITHMS” International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 4, April 2013 ISSN 2319 – 4847

[5] Ref[5]-Arum et al “AN EFFECTIVE SPAM FILTERING FOR DYNAMIC MAIL MANAGEMENT SYSTEM “ ISSN: 2229-6956(ONLINE) ICTACT JOURNAL ON SOFT COMPUTING, APRIL 2012, VOLUME: 02, ISSUE: 03 325

[6] Ref[6]- Arun et al “AN EFFECTIVE SPAM FILTERING FOR DYNAMIC MAIL MANAGEMENT SYSTEM” ISSN: 2229-6956(ONLINE) ICTACT JOURNAL ON SOFT COMPUTING, APRIL 2012, VOLUME: 02, ISSUE: 03 325

[7] Ref[7]- R. Deepa Lakshmi et al. “Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools” (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2760-2766

[8] Ref[8]- Megha et al “Spam Mail Detection through Data Mining – A Comparative Performance Analysis “ I.J. Modern Education and Computer Science, 2013, 12, 31-39 Published Online December 2013 in ECS (http://www.mecs-press.org/) DOI: 10.5815/ijmecs.2013.12.05

[9] Ref[9]- Rajiv Mahajan et al. “Effect of User-Unknown Email addresses in spammers'”/ (IJCSIT) International Journal of Computer Science and Information Tec